## What is claimed is:

| 1  | 1.                | A system for providing capitalization correction for unstructured     |
|----|-------------------|---|
| 2  | excerpts, comp    | orising:  |
| 3  | a prepr           | ocessor to tokenize an excerpt of unstructured content into a set of  |
| 4  | words; and        |   |
| 5  | a capita          | alizer to analyze the set of words for correct capitalization,        |
| 6  | comprising:       |   |
| 7  |                   | an evaluator to evaluate individual characters constituting at least  |
| 8  | one such word     | in the set of words; and  |
| 9  |                   | a filter to skip the at least one such word if determined to be of a  |
| 10 | predefined typ    | e   |
| 1  | 2.                | A system according to Claim 1, further comprising:                    |
| 2  | a title c         | apitalizer to provide one or more of the words with an initial letter |
| 3  | in uppercase a    | nd each remaining letter in lowercase.                                |
| 1  | 3.                | A contain a condition to Chaire 1 C at                                |
| 2  |                   | A system according to Claim 1, further comprising:                    |
| 3  | . *               | nce capitalizer to provide only an initial such word with an initial  |
| 3  | ietter in upperc  | case and each remaining letter in lowercase.                          |
| 1  | 4.                | A system according to Claim 1, further comprising:                    |
| 2  | a word            | analyzer to skip at least one of each such word comprising a          |
| 3  | number, each s    | such word including no vowels, and each such word not occurring       |
| 4  | at a start of a p | hrase and constituting at least one of an article, conjunction,       |
| 5  | preposition.      |   |
| 1  | 5.                | A system according to Claim <sup>-</sup> 1, further comprising:       |
| 2  | a lexico          | on comprising one or more reference words with at least one           |
| 3  | reference word    | defining a form of capitalization for the reference word;             |
| 4  | a match           | ner to match the at least one such word against the reference words,  |
| 5  | the evaluator s   | kipping each such word if a matching reference word is found.         |
| 1  | 6.                | A system according to Claim 1, further comprising:                    |

| 3 | word compri   | ising a noun with no vowels in uppercase.                              |
|---|---|--|
| 1 | 7.  | A system according to Claim 1, further comprising:                     |
| 2 | a tok   | enizer to tokenize the excerpt into the one or more words and one or   |
| 3 | more punctu   | ation marks.   |
| 1 | 8.  | A method for providing capitalization correction for unstructured      |
| 2 | excerpts, con   | nprising:  |
| 3 | toker   | nizing an excerpt of unstructured content into a set of words; and     |
| 4 | analy   | zing the set of words for correct capitalization, comprising:          |
| 5 |   | evaluating individual characters constituting at least one such word   |
| 6 | in the set of   | words; and   |
| 7 |   | skipping the at least one such word if determined to be of a           |
| 8 | predefined ty   | ype.   |
| 1 | 9.  | A method according to Claim 8, further comprising:                     |
| 2 | provi   | ding one or more of the words with an initial letter in uppercase and  |
| 3 | each remaini  | ng letter in lowercase.  |
| 1 | 10.   | A method according to Claim 8, further comprising:                     |
| 2 | provi   | ding only an initial such word with an initial letter in uppercase and |
| 3 | each remaini  | ng letter in lowercase.  |
| 1 | 11.   | A method according to Claim 8, further comprising:                     |
| 2 | skipp   | oing at least one of each such word comprising a number, each such     |
| 3 | word including no vowels, and each such word not occurring at a start of a phrase |  |
| 4 | and constitut   | ting at least one of an article, conjunction, preposition.             |
| 1 | 12.   | A method according to Claim 8, further comprising:                     |
| 2 | main  | taining a lexicon comprising one or more reference words with at       |
| 3 | least one refe  | erence word defining a form of capitalization for the reference word;  |
| 4 | matc  | hing the at least one such word against the reference words; and       |
| 5 | skipp   | oing each such word if a matching reference word is found.             |

a proper noun capitalizer to provide the individual letters in each such

2

0365.US.UTL.ap5 - 18 -

| 1   | 13.             | A method according to Claim 8, further comprising:                   |
|-----|-----------------|--|
| 2   | provid          | ing the individual letters in each such word comprising a noun with  |
| 3   | no vowels in u  | ippercase.   |
| 1 . | . 14            | A mostly of according to Oleine O. forther according                 |
| 1   | 14.             | A method according to Claim 8, further comprising:                   |
| 2   |                 | zing the excerpt into the one or more words and one or more          |
| 3   | punctuation m   | arks.  |
| 1   | 15.             | A computer-readable storage medium holding code for performing       |
| 2   | the method acc  | cording to Claim 8.  |
| 1   | 16.             | An apparatus for providing capitalization correction for             |
| 2   | unstructured e  | xcerpts, comprising:   |
| 3   | means           | for tokenizing an excerpt of unstructured content into a set of      |
| 4   | words; and      |  |
| 5   | means           | for analyzing the set of words for correct capitalization,           |
| 6   | comprising:     |  |
| 7   |                 | means for evaluating individual characters constituting at least one |
| 8   | such word in t  | he set of words; and   |
| 9   |                 | means for skipping the at least one such word if determined to be    |
| 0   | of a predefined | d type.  |
| 1   | 17.             | A system for building a lexicon for use in capitalization correction |
| 2   | for unstructure | ed excerpts, comprising:   |
| 3   | a rippe         | r assembling a list of word sets from unstructured content, each     |
| 4   | word set comp   | orising a word and at least one variation on capitalization;         |
| 5   | an agg          | regator aggregating each word set, comprising:                       |
| 6   |                 | an analyzer identifying at least one word set comprising significant |
| 7   | statistics; and |  |
| 8   |                 | a non-standard capitalization selector selecting at least one such   |
| 9   | variation withi | in the identified word set having a non-standard capitalization, and |
| Λ   | adding the at 1 | east and such varieties to the lavides                               |

| 1 | 18. A system according to Claim 17, further comprising:                            |
|---|--|
| 2 | a tokenizer tokenizing the excerpt into the one or more words and one or           |
| 3 | more punctuation marks.  |
| 1 | 19. A system according to Claim 18, wherein hyphenated words are                   |
| 2 | split into a plurality of the words.   |
| 1 | 20. A system according to Claim 17, wherein at least one variation                 |
| 2 | appearing at the start of a sentence is skipped.                                   |
| 1 | 21. A system according to Claim 20, wherein the non-standard                       |
| 2 | capitalization comprises the at least one variation occurring in an excerpt having |
| 3 | fewer than half of individual letters provided in uppercase.                       |
| 1 | 22. A system according to Claim 17, further comprising:                            |
| 2 | a normalizer normalizing a plurality of the words extracted relative to a          |
| 3 | source of the structured Web content.  |
| 1 | 23. A system according to Claim 17, wherein the significant statistic              |
| 2 | comprises at least four occurrences of at least one such variation within a word   |
| 3 | set.   |
| 1 | 24. A system according to Claim 17, wherein the non-standard                       |
| 2 | capitalization comprises the at least one variation having any individual letter   |
| 3 | other than the first individual letter provided in uppercase.                      |
| 1 | 25. A system according to Claim 17, further comprising:                            |
| 2 | a standard capitalization selector selecting at least one such variation           |
| 3 | within the identified word set having a standard capitalization, and adding the a  |
| 4 | least one such variation to the lexicon.   |
| 1 | 26. A system according to Claim 17, further comprising:                            |
| 2 | a validator applying implicit rules for capitalization, and skipping each a        |
| 3 | least one variation subject to at least one such implicit rule.                    |

0365.US.UTL.ap5 - 20 -

| 1   | 27. A system according to Claim 26, wherein the implicit rules                        |
|-----|---|
| 2   | comprise skipping at least one variation based on position within a sentence or       |
| 3   | phrase.   |
| 1   | 28. A system according to Claim 26, wherein the implicit rules                        |
| 2   | comprise at least one of a number, having no vowels, and constituting at least one    |
| 3 . | of an article, conjunction and preposition.   |
| 1   | 29. A system according to Claim 26, wherein the implicit rules                        |
| 2   | comprise normalizing a number of occurrences for each at least one variation          |
| 3   | using at least one of a normalizing function and relative to a source of the at least |
| 4   | one variation.  |
| 1   | 30. A system according to Claim 26, wherein the implicit rules                        |
| 2   | comprise accommodating multiple forms of capitalization for each at least one         |
| 3   | variation by annotating each capitalization form with a frequency count and           |
| 4   | skipping those of the each at least one variation occurring infrequently.             |
| 1   | 31. A system according to Claim 17, further comprising:                               |
| 2   | a hash table maintaining the lexicon.   |
| 1   | 32. A system according to Claim 31, further comprising:                               |
| 2   | at least one record specifying at least one such word as a key into the hash          |
| 3   | table, and associating at least one such variation within the word set as a preferred |
| 4   | capitalization.   |
| 1   | 33. A method for building a lexicon for use in capitalization correction              |
| 2   | for unstructured excerpts, comprising:  |
| 3   | assembling a list of word sets from unstructured content, each word set               |
| 4   | comprising a word and at least one variation on capitalization;                       |
| 5   | aggregating each word set, comprising:  |
| 6   | identifying at least one word set comprising significant statistics:                  |

| 7   |                | selecting at least one such variation within the identified word set    |
|-----|----------------|---|
| 8   | having a non   | -standard capitalization; and   |
| 9.  |                | adding the at least one such variation to the lexicon.                  |
| 1   | 34.            | A method according to Claim 33, further comprising:                     |
| 2   | token          | izing the excerpt into the one or more words and one or more            |
| 3   | punctuation i  | marks.  |
| 1   | 35.            | A method according to Claim 34, further comprising:                     |
| 2   | splitti        | ng hyphenated words into a plurality of the words.                      |
| 1   | 36.            | A method according to Claim 33, further comprising:                     |
| 2   | skipp          | ing at least one variation which may be at the start of a sentence.     |
| 1   | 37.            | A method according to Claim 36, wherein the non-standard                |
| 2   | capitalization | a comprises the at least one variation occurring in an excerpt having   |
| 3   | fewer than ha  | alf of individual letters provided in uppercase.                        |
| 1   | 38.            | A method according to Claim 33, further comprising:                     |
| 2   | norm           | alizing a plurality of the words extracted relative to a source of the  |
| 3   | structured W   | eb content.   |
| 1   | 39.            | A method according to Claim 33, wherein the significant statistics      |
| 2   | comprises at   | least four occurrences of at least one such variation within a word     |
| 3   | set.           | · ·   |
| 1 . | 40.            | A method according to Claim 33, wherein the non-standard                |
| 2   | capitalization | comprises the at least one variation having any individual letter       |
| 3   | other than the | e first individual letter provided in uppercase.                        |
| 1   | 41.            | A method according to Claim 33, further comprising:                     |
| 2   | select         | ing at least one such variation within the identified word set having a |
| 3.  | standard capi  | talization, and adding the at least one such variation to the lexicon.  |
| 1   | 42.            | A method according to Claim 33, further comprising:                     |

| 2 | applying implicit rules for capitalization; and                                       |
|---|---|
| 3 | skipping each at least one variation subject to at least one such implicit            |
| 4 | rule.   |
| 1 | 43. A method according to Claim 42, wherein the implicit rules                        |
| 2 | comprise skipping at least one variation based on position within a sentence or       |
| 3 | phrase.   |
| 1 | 44. A method according to Claim 42, wherein the implicit rules                        |
| 2 | comprise at least one of a number, having no vowels, and constituting at least one    |
| 3 | of an article, conjunction and preposition.   |
| 1 | 45. A method according to Claim 42, wherein the implicit rules                        |
| 2 | comprise normalizing a number of occurrences for each at least one variation          |
| 3 | using at least one of a normalizing function and relative to a source of the at least |
| 4 | one variation.  |
| 1 | 46. A method according to Claim 42, wherein the implicit rules                        |
| 2 | comprise accommodating multiple forms of capitalization for each at least one         |
| 3 | variation by annotating each capitalization form with a frequency count and           |
| 4 | skipping those of the each at least one variation occurring infrequently.             |
| 1 | 47. A method according to Claim 33, further comprising:                               |
| 2 | maintaining the lexicon structured as a hash table.                                   |
| 1 | 48. A method according to Claim 47, further comprising:                               |
| 2 | specifying at least one such word as a key into the hash table; and                   |
| 3 | associating at least one such variation within the word set as a preferred            |
| 4 | capitalization.   |
| 1 | 49. A computer-readable storage medium holding code for performing                    |
| 2 | the method according to Claim 33.   |
| 1 | 50. An apparatus for building a lexicon for use in capitalization                     |

correction for unstructured excerpts, comprising:

| 3  | means for assembling a list of word sets from unstructured content, each |
|----|--|
| 4  | word set comprising a word and at least one variation on capitalization; |
| 5  | means for aggregating each word set, comprising:                         |
| 6  | means for identifying each word set comprising significant               |
| 7  | statistics;  |
| 8  | means for selecting at least one such variation within the identified    |
| 9  | word set having a non-standard capitalization; and                       |
| 10 | means for adding the at least one such variation to the lexicon.         |